

Вероятность переобучения плотных и разреженных семейств алгоритмов

И. О. Толстихин

iliya.tolstikhin@gmail.com

Научный руководитель — к.ф.-м.н. К. В. Воронцов
октябрь 2010

Содержание

1 Проблема переобучения

- Постановка задачи
- Завышенность классических оценок
- Экспериментальное измерение факторов завышенности

2 Расслоенные и связные семейства алгоритмов

- Модельные семейства
- Точная оценка вероятности переобучения для шара
- Приближение оценки шара d его нижними слоями

3 Плотные семейства алгоритмов

- Модельное семейство
- Точная оценка вероятности переобучения для слоя шара
- Разреженные подмножества слоя шара

Постановка задачи

Объекты $\mathbb{X} = \{x_1, \dots, x_L\}$; алгоритмы $A = \{a_1, \dots, a_D\}$;

$I(a, x) = [\text{алгоритм } a \text{ ошибается на объекте } x]$;

$n(a, X)$ — число ошибок a на выборке X ;

$\nu(a, X) = n(a, X)/|X|$ — частота ошибок a на выборке X .

Статистическое обучение: $a(X) = \arg \min_{a \in A} n(a, X)$.

Переобучение: $\nu(a, X') - \nu(a, X) \geq \varepsilon$.

Пример. Бинарная $L \times D$ -матрица ошибок, $L = \ell + k$:

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_D	
x_1	1	1	0	0	0	1	1	1	1	X , обучение
x_2	0	0	0	0	1	1	1	1	1	
x_ℓ	0	0	1	0	0	0	0	0	0	
x'_1	0	0	0	1	1	1	1	1	0	X' , контроль
x'_2	0	0	0	1	0	0	0	1	1	
x'_k	0	1	1	1	1	1	0	0	0	

Задача: оценить вероятность переобучения.

Основная задача — оценить вероятность переобучения

$$Q_\varepsilon = P\left[\nu(a(X), X') - \nu(a(X), X) \geq \varepsilon\right].$$

Теорема (Вапник и Червоненкис, 1971)

Для любой выборки, метода обучения и числа $\varepsilon \in (0, 1)$

$$Q_\varepsilon \leq P\left[\sup_{a \in A} (\nu(a, X') - \nu(a, X)) \geq \varepsilon\right] \leq |A| \max_m H_L^{\ell, m} \left(\frac{\ell}{L} (m - \varepsilon k)\right),$$

где $H_L^{\ell, m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ — гипергеометрическая функция распределения.

Основная проблема — завышенность оценки:

в реальных задачах $|A| \sim 10^6 - 10^{12}$

Выявление причин завышенности

Основные причины завышенности:

- не учитывается *расслоение* семейства алгоритмов:
чем выше уровень ошибок m , тем меньше вероятность получить алгоритм в результате обучения (завышенность в 10^2 – 10^5 раз);
- не учитывается *связность* семейства алгоритмов:
чем больше схожих алгоритмов, тем сильнее завышенность (завышенность в 10^3 – 10^4 раз).

Реальные семейства, как правило, расслоены и связны.

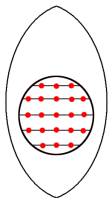
Vorontsov K. V. Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. MAIK Nauka. No 2, Vol. 18, 2008, Pp. 243–259.

Гипотеза

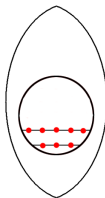
Вероятность переобучения расслоенного и связного множества алгоритмов может быть аппроксимирована вероятностью переобучения его подмножества, состоящего из существенно различных алгоритмов нижних слоев.

Модельные семейства

A_m — m -й слой булева куба $\{0, 1\}^L$ — множество алгоритмов, допускающих на полной выборке m ошибок.



Шар алгоритмов



Нижние слои шара

Точная оценка для шара алгоритмов

A — хэммингов шар радиуса r_0 с центром в некотором элементе m -го слоя.

Теорема (Оценка для шара)

Точная оценка вероятности переобучения этого семейства есть

$$Q_\varepsilon = \sum_{i=0}^{r_0} h_L^{\ell, m}(i) \frac{\sum_{r_1=0}^{r_0} \sum_{n_1=0}^{r_1} S(n_1, r_1, i) [m + r_1 - 2n_1 \geq \varepsilon k]}{\sum_{r_2=0}^{r_0} \sum_{n_2=0}^{r_2} S(n_2, r_2, i)} + \\ + \sum_{i=r_0+1}^{\lfloor s_d + \frac{rk}{L} \rfloor} h_L^{\ell, m}(i).$$

где $h_L^{\ell, m}(i) = \frac{C_m^i C_{L-m}^{\ell-i}}{C_L^\ell}$, $S(n, r, i) = C_{m-i}^{n-i} C_{k-m+i}^{r-n}$, $s_d = \frac{\ell}{L}(m - \varepsilon k)$.

Вклады слоёв шара

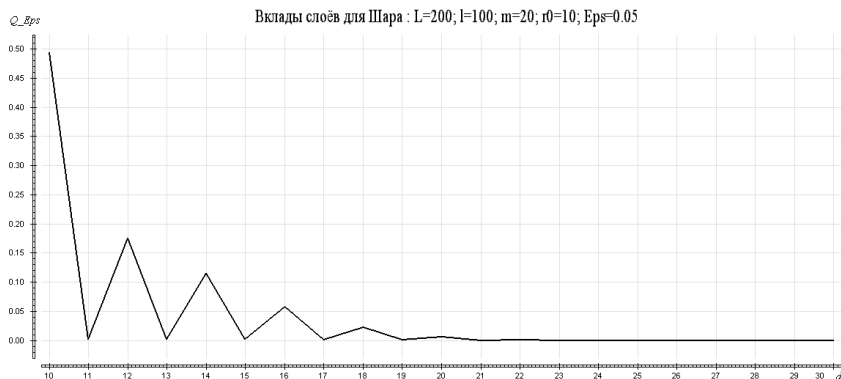


Рис.: Вклады слоёв шара в вероятность переобучения

Точная оценка для нижних слоев шара

A — хэммингов шар радиуса r_0 с центром в некотором элементе m -го слоя в пересечении с d его нижними слоями.

Теорема (Оценка для нижних слоев шара)

Точная оценка вероятности переобучения этого семейства есть

$$Q_\varepsilon = \sum_{i=0}^{r_0} h_L^{\ell,m}(i) \frac{\sum_{r=0}^{r_0} \sum_{n=0}^r S'(n, r, i) [m + r - 2n \geq \varepsilon k]}{\sum_{r=0}^{r_0} \sum_{n=0}^r S'(n, r, i)} + \\ + \sum_{i=r_0+1}^{\lfloor s_d + \frac{rk}{L} \rfloor} h_L^{\ell,m}(i),$$

где

$$h_L^{\ell,m}(i) = \frac{C_m^i C_{L-m}^{\ell-i}}{C_L^\ell}, \quad S'(n, r, i) = C_{m-i}^{n-i} C_{k-m+i}^{r-n} [r + r_0 + 1 \leq 2n + d], \\ s_d = \frac{\ell}{L}(m - \varepsilon k).$$

Оценки для нижних слоёв шара

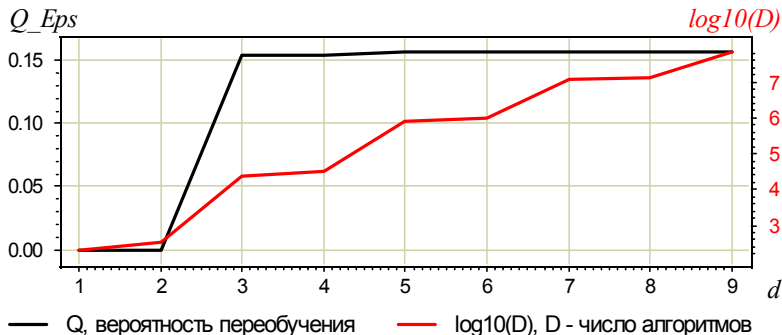


Рис.: Зависимость Q_ϵ и $\log_{10} |A|$ от числа d нижних слоёв шара, при $\ell = k = 100$, $m = 10$, $r_0 = 4$, $\epsilon = 0.07$.

Зависимость Q_ϵ от числа алгоритмов в семейства

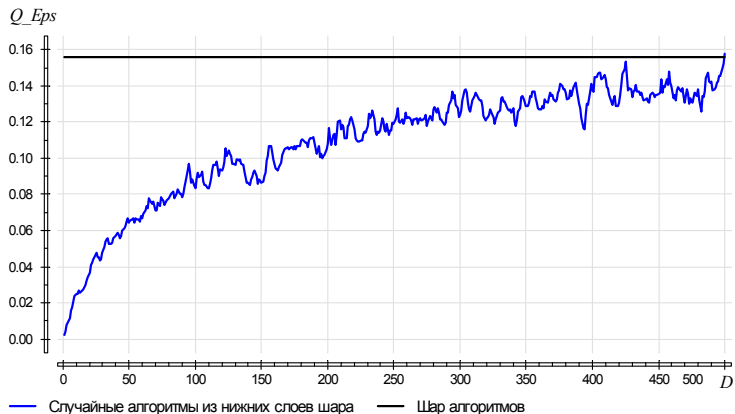
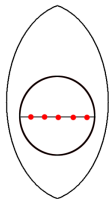


Рис.: Оценки Q_ϵ для шара алгоритмов и D случайных алгоритмов из трех его нижних слоев, при $\ell = k = 100$, $m = 10$, $r_0 = 4$, $\epsilon = 0.07$.

Модельное семейство



Слой шара

Точная оценка для слоя шара

$B(m, r_0)$ — пересечение m -го слоя булева куба $\{0, 1\}^L$ с хэмминговым шаром радиуса r_0 с центром в некотором элементе m -го слоя.

Теорема (Оценка для слоя шара)

Если μ минимизирует частоту ошибок на обучающей выборке, то достижимая верхняя оценка вероятности переобучения этого семейства есть

$$Q_\varepsilon = H_L^{\ell, m}(s_d + \lfloor r_0/2 \rfloor),$$

где $s_d = \frac{\ell}{L}(m - \varepsilon k)$.

Зависимость Q_ϵ от радиуса шара

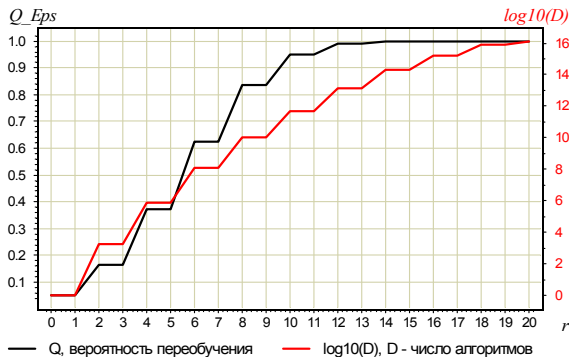


Рис.: Зависимость вероятности переобучения Q_ϵ и $\log_{10} |A|$ от радиуса шара r_0 для m -го слоя шара, при $l = k = 100$, $m = 10$, $\epsilon = 0,05$.

Зависимость Q_ϵ от числа алгоритмов в семейства

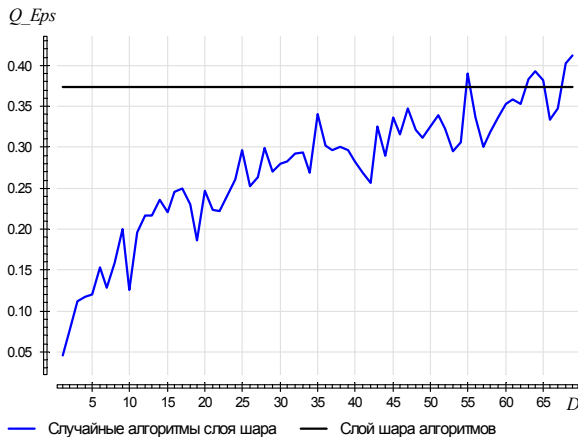


Рис.: Оценки Q_ϵ для слоя шара и его случайного подмножества из D алгоритмов, при $l = k = 100$, $m = 10$, $r_0 = 4$, $\epsilon = 0,05$.

Лемма

Пусть $A \subset A_m$ и $a \in A_m \setminus A$.

Тогда $Q_\varepsilon(A) \leq Q_\varepsilon(A \cup \{a\})$.

$S(m, r_0)$ — пересечение m -го слоя булева куба $\{0, 1\}^L$ с хэмминговой сферой радиуса r_0 с центром в некотором элементе m -го слоя.

Теорема

Пусть $m = n(a, \mathbb{X})$ и $k \geq m + \lfloor r_0/2 \rfloor$.

Тогда вероятности переобучения множеств $B(m, r_0)$ и $S(m, 2\lfloor r_0/2 \rfloor)$ совпадают.

Подмножества B' и B''

$$B'(m, r_0), B''(m, r_0) \subset S(m, 2\lfloor r_0/2 \rfloor);$$

$$|B'(m, r_0)| = C_m^{\lfloor r_0/2 \rfloor};$$

$$|B''(m, r_0)| = C_m^{\lfloor r_0/2 \rfloor} \frac{L-m}{\lfloor r_0/2 \rfloor};$$

Оценки вероятности переобучения для подмножеств B' и B''

Теорема (Подмножество B')

Точная оценка вероятности переобучения семейства $B'(m, r_0)$ есть

$$Q_\varepsilon = \sum_{i=0}^m \sum_{j=0}^h \sum_{p=0}^{\delta} \frac{C_m^i C_h^j C_\delta^p}{C_L^\ell} \times \\ \times \left([i < \delta] [p \leq s_d(\varepsilon)] + [i \geq \delta] [i + p \leq \delta + s_d(\varepsilon)] \right),$$

где $h = L - m - \delta$, $s_d(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$, $\delta = \lfloor r_0/2 \rfloor$.

Теорема (Подмножество B'')

Пусть $\ell < \frac{L-m}{\lfloor r_0/2 \rfloor}$. Тогда вероятности переобучения множеств $B''(m, r_0)$ и $B(m, r_0)$ совпадают.

Приближение оценки слоя шара подмножествами B' и B''

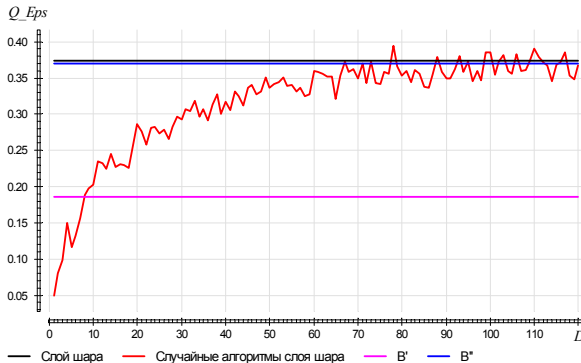


Рис.: Оценки Q_ε для слоя шара, множеств $B'(m, r_0)$, $B''(m, r_0)$ и случайного подмножества D алгоритмов из слоя шара, при $l = k = 100$, $m = 10$, $r_0 = 4$, $\varepsilon = 0,05$.

Результаты и открытые вопросы

Полученные результаты.

- Показано, что вероятность переобучения расслоенных и связанных семейств алгоритмов может быть аппроксимирована вероятностью переобучения его подмножества, состоящего из существенно различных алгоритмов, допускающих мало ошибок.
- Показано, что вероятность переобучения плотного семейства алгоритмов может быть аппроксимирована его немногочисленным разреженным подмножеством.

Открытые вопросы.

- Вопрос о возможности аппроксимации вероятности переобучения реальных семейств алгоритмами из их нижних слоев остаётся открытым.
- Эксперименты показывают, что выбранные разреженные подмножества избыточны.